

EDUCATION

Boston University, Boston, MA Ph.D. in Economics Field of interest: Macroeconomics, Econometrics, Machine Learning	2019.09 – 2025.08
Renmin University of China, Beijing, China BS degree, double major: Mathematics and Finance	2015.09 – 2019.06 GPA: 3.82/4

SKILLS

Programming & Tools: Python (PyTorch, Scikit-learn, NumPy, Pandas), SQL, R, Julia, Spark, Git, Linux
Machine Learning: Deep Learning, Random Forest, XGBoost, SVM, Recommender Systems, NLP, Clustering
Data Science & Analytics: A/B Testing, Causal Inference, Bayesian Methods, MCMC, Time Series
Big Data & Deployment: FastAPI, Gradio, FAISS, AWS, Docker

WORK AND RESEARCH EXPERIENCE

Graduate Researcher in Causal Inference and Monetary Policy, <i>Boston University</i> • Applied K-Means Clustering on 10+ years of financial data to segment households based on asset allocation, incorporating demographic attributes such as age, income, and wealth distribution • Leveraged causal inference techniques (DID) to quantify the impact of QE on household asset allocation, enabling data-driven policy recommendations. • Developed a Heterogeneous Agent New Keynesian (HANK) model and calibrated it using Simulated Method of Moments (SMM) to model macro-financial dynamics post-QE • Conducted large-scale quantitative analysis using Python (Pandas, NumPy, StatsModels), demonstrating that redistribution amplifies short-term QE effects by 6% and dampens long-run effects by ~2%	2021.09 – 2025.05
Research Assistant in Applied Machine Learning and Financial Markets, <i>Boston University</i> • Built predictive models using Linear Regression and XGBoost to analyze the relationship between investment decisions, market sentiment, and stock price movements, identifying key factors contributing to stock overpricing • Developed a two-agent simulation model to evaluate information asymmetry, managerial compensation structures, and investment behavior, optimizing model estimation via Simulated Method of Moments (SMM) • Conducted large-scale financial data analysis using Python (Pandas, Scikit-learn, NumPy), quantifying the total social welfare loss from stock overpricing at ~2% • Engineered feature selection pipelines and applied hyperparameter tuning to improve XGBoost model accuracy by 15%, ensuring robust financial predictions	2022.09 – 2025.05

PROJECT EXPERIENCE

Multi-modal Product Recommendation System with Large Language Model • Designed a multi-modal recommendation system by adapting LLaVA to generate high-quality text and image embeddings within a shared latent space using a contrastive learning objective • Implemented a cosine similarity-based ranking algorithm to match user queries with product embeddings, achieving 92% average precision and reducing irrelevant recommendations by 10%, leading to improved user engagement • Optimized large-scale vector search using FAISS, enabling real-time, low-latency similarity searches on millions of product embeddings, making the system highly scalable for e-commerce applications • Conducted hyperparameter tuning and model compression to enhance system performance, reducing memory usage by 15% while maintaining high recommendation accuracy • Designed and evaluated A/B testing experiments to assess the system's business impact, demonstrating a 5% increase in user retention through enhanced recommendation relevance	2024.12 - 2025.02
Dialogue Summarization with Large Language Model • Fine-tuned LLaMA-2 on diverse dialogue datasets using supervised learning, optimizing for concise and human-like summaries while improving generalization across varied conversational styles • Engineered an evaluation pipeline to measure summarization quality using ROUGE-L (45.2) and BLEU (37.8) scores, ensuring high coherence and factual consistency • Optimized inference efficiency by applying Low-Rank Adaptation (LoRA), reducing compute requirements by 35% while maintaining high-quality summaries, making the system cost-effective for real-world deployment • Developed and deployed an interactive web-based demo using Gradio, enabling real-time user input and summarization, improving accessibility for non-technical users	2024.09 - 2024.12

Pengyue Zhu

5 Royce Rd, Allston, MA, 02134

857-415-0748 | pyuzhu@bu.edu

• Designed scalable model deployment pipelines for production, integrating with FastAPI and GPU-accelerated inference, reducing response latency for large-scale applications

SELECTED HONORS AND AWARDS

Dean’s Fellowship, Boston University	2019 - 2025
National Scholarship (top 0.2% nationwide), the Ministry of Education, P.R. China	2018
Academic Perfection Recognition, UC Davis	2018
President Scholarship, Renmin University of China	2017